Design of Hybrid Interpretable Predictive Machine Learning Model for Air Pollution Prediction in South Asia Region

Synopsis of Research Proposal

Submitted for the Registration in the

Ph.D. Program (Part-Time) in Science

By

SHWETA AGRAWAL

Under Supervision of

Dr. Rana Mukherji

The ICFAI University, Jaipur

THE ICFAI UNIVERSITY JAIPUR

Abstract

Air pollution poses a significant threat to public health and the environment, particularly in densely populated regions like South Asia. Accurate prediction of air pollution levels is essential for effective air quality management and the protection of human health. In this study, we propose a novel hybrid interpretable predictive machine learning model for air pollution prediction in South Asia.

The dataset used for model training and evaluation will include historical air quality measurements, meteorological variables, and geographical features specific to the South Asian region. Feature engineering techniques will be employed to extract relevant information from the raw data, ensuring maximum information utilization by the model.

To validate the performance of the proposed model, extensive experiments will be conducted on real-world air pollution datasets from various locations in South Asia. The model's predictions will be compared with other state-of-the-art air pollution prediction models, demonstrating its superiority in terms of accuracy and interpretability.

The interpretability of the hybrid model will allow policymakers and stakeholders to comprehend the factors influencing air pollution levels in the region, aiding in the formulation of targeted pollution control strategies. Furthermore, the model can be integrated into existing air quality monitoring systems, providing real-time predictions to enhance public awareness and health preparedness.

The results will indicate that the hybrid interpretable predictive machine learning model offers a promising solution for air pollution prediction in South Asia. By bridging the gap between model accuracy and explainability, it will address the critical challenges associated with traditional black-box machine learning models, making it a valuable tool for sustainable air quality management and environmental protection in the region in the future.

Table of Contents

1. Introduction 4
2. Literature Review 9
3. Present Study 15
3.1 Statement of the problem: 15
3.2 Significance of Study 15
3.3 Objectives of study 16
3.4 Research Questions 16
4. Methodology 17
4.1 Research Design 17
4.2 Sample/ Participants 18
4.3 Data Collection 18
5. Data Analysis / Discussion 18
6. Work Plan of the Research 19
7. References 19

1. Introduction

The invisible gas mixture that surrounds Earth is labeled as air. Most living things depend on essential elements like oxygen and nitrogen found in air to exist. Of course, one of those species is human beings (Homo sapiens). In some cases, "atmosphere" is used in place of "air." Without air no life would exist, animals would not be able to live. There won't be any sound without air.

Air Quality

The degree to which the air is healthy or clean enough for people or the environment is referred to as air quality. The absence of harmful compounds in the atmosphere is a sign of good air quality. Poor air quality can aggravate asthma and other respiratory disorders, irritate the eyes, nose, and throat, induce shortness of breath, and have negative effects on the heart and cardiovascular system. Long-term exposure to polluted air can result in more severe issues. There is growing proof that breathing dirty air is also hazardous for your brain. Over the past ten years, researchers have discovered that excessive air pollution may harm children's cognitive development, raise the risk of cognitive ageing in adults, and maybe even cause depression. Children who did not previously have any risk factors for mental health issues may experience anxiety and depression symptoms or develop them as a result of exposure to polluted air. When the Quality of air is clean, it is clear and only includes minute quantities of chemical and solid impurities. High levels of pollutants in poor air are typically foggy, harmful to human health, and hazardous to the ecosystem. The concentration of pollutants in the air at a certain location is the basis for the Air Quality Index (AQI), which is used to describe air quality.

Humans Affect Air Quality

Geographical features of a place, such mountain ranges, coastlines, and terrain that has been transformed by humans, can influence where air pollutants are concentrated or spread. However, the types and amounts of contaminants that enter the air have a much higher impact on the quality of the air. While some air pollutants are created by natural events like volcanic eruptions and dust storms, most are brought about by human activity. Exhaust from automobiles, smoke from coal-burning power plants, and toxic gases from industries are a few examples of man-made air pollution.

Winds Affect Air Quality

Wind patterns have an impact on AQ since winds are responsible for transporting air pollution. For example, a coastal area with an inland mountain range may have higher levels of air pollution during the day as sea breezes bring pollutants over the land, and lower levels during the evening when the breeze's direction shifts and carries pollutants out over the ocean.

Temperature Affects Air Quality

Temperature can have an effect on AQ as well. Urban places' air quality frequently gets worse throughout the winter. Exhaust emissions may become trapped at the surface under a layer of thick, frigid air at lower air temperatures. During the summer, when hot air rises, pollutants from the Earth's surface are spread across the upper troposphere. However, greater sunshine causes harmful ground-level ozone to increase.

Air Pollution

Air pollution is caused by the atmospheric release of harmful suspended particles. In the majority of densely populated places, it is a big problem. Because of air pollution, more deaths are reported and new diseases are found every year. Several causes, such as industrialization and globalisation, are contributing to an increase in the rate of air pollution. Around the world, many monitoring stations have been installed to keep a check on the air quality. The problem of air pollution is getting worse due to a number of factors, including household warming, increased car traffic, energy production, rising commercial and industrial activity, and urban population development. Because air quality has an impact on health, it is critical to monitor and set air quality standards in metropolitan areas.

Air Quality Index (AQI)

The AQI is a metric for daily reporting on AQ. It measures the rate at which air pollution impairs a person's health. To help people better understand how their local AQ affects their health, the AQI was developed. The Environmental Protection Agency (EPA) determines the AQI for five principal air pollutants for which national air quality regulations have been created to protect human health.

1. Ground-level ozone

2. Particle pollution/particulate matter (PM2.5/pm 10)

3. Carbon Monoxide

4. Sulfur dioxide

5. Nitrogen dioxide

For the layperson to evaluate the local air quality, the MoEFCC defined the Air Quality Index (AQI) as "One Number- One Colour- One Description."

An indicator of how poisonous the air is now and might become in the near future is the AQI. It also alerts you to any possible detrimental health effects. It largely focuses on the detrimental effects on your health that could occur hours or days after breathing contaminated air.

To make it easier for people to comprehend the current level of the air quality, a scale known as the AQI was developed. The complex data regarding the many contaminants and toxins in the air is converted into a numerical value (index value), a phrase, and a colour.

The main objective of the AQI is to inform and forewarn the public about the risks of exposure to background pollution levels. When the AQI increases, a sizable section of the populace is exposed to substantial health risks.

The degree of air pollution and the associated health risks increase as the AQI value rises. For more than three decades, several developed nations have adopted the idea of AQI extensively. Real-time air quality data is promptly shared by AQI.

AQI Calculation

For reporting AQI, different nations have various point systems. A rating between 0 and 50 on a 500 point scale, for instance, is considered good in the United States. Ratings in the dangerous range of 301 to 500 are considered. India too uses the 500-point system. Daily concentrations of the main pollutants are recorded via monitors. Using standardised formulas created by the EPA, these raw readings are transformed into a separate AQI value for each pollutant (ground-level ozone, particle pollution, carbon monoxide, and sulphur dioxide). The AQI value for that day is reported as the highest of these AQI values.

Air Quality Index Categories

Good (0–50) - Minimal Impact

Satisfactory (51–100) - May cause minor breathing difficulties in sensitive people.

Moderately polluted (101–200) - May cause breathing difficulties in people with lung disease like asthma, and discomfort to people with heart disease, children and older adults.

Poor (201–300) - May cause breathing difficulties in people on prolonged exposure, and discomfort to people with heart disease

Very Poor (301–400) - May cause respiratory illness in people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases.

Severe (401-500) - May cause respiratory issues in healthy people, and serious health issues in people with lung/heart disease. Difficulties may be experienced even during light physical activity.

Importance of AQI

It is essential for those suffering from illnesses caused by exposure to air pollution to be aware of daily levels of air pollution. As well as measuring air quality, the AQI provides suggestions for improving it. At various levels of air quality, this index provides specific advice on how to protect oneself against various health hazards posed by air pollution.

People who work or exercise outside frequently are especially vulnerable to air pollution and may experience respiratory problems when the pollution is severe. Since most people in developing countries are unfamiliar with technical terminology and units of measurement (such as ppm/ppb/ or g/mg3), AQI is even more crucial. With each number and colour representing a different health risk category, the AQI makes air quality easier to understand by decoding it into unambiguous numbers and colors.

Doctors can also use the AQI to forecast when respiratory distress cases increase whenever the AQI is high, since patients with respiratory distress are susceptible when the AQI is high. This report contains a large number of pollution-ridden cities and countries located in the South Asian region. Four of the five most polluted cities are located in this region most polluted countries.

Out of 147 cities with monitoring data in this region in 2019, only one (Sanandaj, Iran) achieved WHO PM2.5 target values. Even though the region's pollution sources vary, frequent sources include industrial, open agricultural burning, coal combustion, domestic biomass burning, and transportation emissions. In many cities in the area, especially in Pakistan and India, PM2.5 levels rose from 2018 to 2019. In cities with comparable PM2.5 data from 2018 and 2019, this led to a regional reduction in PM2.5 levels of 14.8%. Increased monitoring data, a slowing economy, favourable weather circumstances, and government action can all be held responsible for a large portion of this. India's National Clean Air Program launched in 2019, outlining new tactics for achieving the program's lofty PM2.5 targets. [2019 World Air Quality Report]

2. Literature Review

Because contaminants and particles are volatile, dynamic, and highly variable in both time and place, predicting air quality is a challenging endeavour. The observed significant impact of air pollution on people's health and the environment is making it increasingly important to model, predict, and monitor air quality, particularly in urban areas. In the study, Researcher has used support vector regression (SVR), a well-liked machine learning technique, to forecast pollutant and particle levels and to estimate the air quality index (AQI). Radial basis function (RBF) was the kind of kernel among the investigated choices that gave SVR the most precise predictions. It was shown that employing all of the available variables was more effective than using principal component analysis to choose features. The results show that SVR with RBF kernel enables to precisely forecast hourly pollutant concentrations, such as carbon monoxide, sulphur dioxide, nitrogen dioxide, ground-level ozone, and particulate matter 2.5, as well as the hourly AQI for the state of California. On unseen validation data, classification into the six AQI categories set forth by the US Environmental Protection Agency was carried out with an accuracy of 94.1%. (Castelli M, 2020)

Since air quality is increasingly affecting people's health, the government must take required steps to forecast it. The air's quality is measured by the air quality index. Carbon dioxide, nitrogen dioxide, and carbon monoxide are among the air pollutants that cause air pollution and are released when natural gas, coal, and wood are burned, as well as by industry, cars, and other sources. Serious illnesses including lung cancer, brain sickness, and even death can be brought on by air pollution. The air quality index is determined with the aid of machine learning techniques. Numerous studies are being conducted in this area, however the outcomes are still not reliable. Datasets from Kaggle and air quality monitoring stations are accessible and separated into Training and Testing. Linear Regression, Decision Tree, Random Forest, Artificial Neural Network, and Support Vector Machine are some of the machine learning techniques used for this (Madan T, 2020).

The achievement of high-accuracy air quality index prediction is of tremendous research value in the context of carbon neutrality and air pollution control. Data from January 2014 to December 2019 and data from January 2020 to December 2021 are used as the training set and test set,

respectively, in this paper, which uses Beijing as its study area. For analysis and prediction, in this paper CEEMDAN-ARMA-LSTM model is used. To increase the exploitation of data and information, the CEEMDAN model is utilised to break down the data. The ARMA model receives the smooth non-white noise components, and the LSTM model receives the remaining components and residuals. The outcomes demonstrate that this model has the minimum MAE, MAPE, MSE, and RMSE. MAE improved by 22.5%, 53.4%, and 21.5% compared to the CEEMDAN-LSTM, LSTM, and ARMA-GARCH models, while MAPE by 21.4%, 55.3%, and 26.1%, MSE by 39.9%, 76.9%, and 28.5%, and RMSE by 22.5%, 52.0%, and 15.4%. The accuracy increase is substantial and offers promising application possibilities (Sun Y, 2022).

In this paper, author applied a sensor that observed air quality big data, information theory-based predictor significance identification, and PEK-based machine learning to air quality index (AQI) analysis and prediction in order to investigate which factors most affect air quality and to provide a useful tool to assist the prediction and early warning of air pollution in urban areas. Author discovered that there is a strong correlation between air quality stability and absolute air quality, and that air quality improvement can also enhance stability. Cities in the south and west have superior air quality than those in the north and east. AQI time series of cities with closer geophysical locations have a closer relationship with others. PM2.5, PM10, and SO2 are the most important impact factors. The machine learning-based prediction is useful for AQI prediction and early warning. (Chen S, 2018)

The researcher investigated various big-data and machine learning based techniques for air quality forecasting. This paper reviewed the published research results relating to air quality evaluation using methods of artificial intelligence, decision trees, deep learning etc many researchers began to use the big data analytics approach due to advancements in big data applications and availability of environmental sensing networks and sensor data (Kang GK, 2018).

K-nearest neighbour is a traditional data mining approach. This function classifies a new instance by using the class of the k nearest neighbours. One of the several mathematical distance metrics is used to compute the distance. In order to forecast the value of the air quality index, the technique is utilised in this paper's air quality forecast domain. This index is intended to classify the degree of pollution and to alert the public to particular potential instances of pollution (EG, 2010).

This research has integrated principal component analysis (PCA), a genetic algorithm (GA) and a backpropagation (BP) neural network to evaluate the regional disaster risk. Hazard risk, hazard-laden environment sensitivity, hazard-bearing body vulnerability and disaster resilience were used to measure the degree of disaster risk. The main findings were: (1) the air pollution disaster risk index of Liaoning Province, Beijing, Shanghai and Guangdong Province increased year by year from 2010 to 2019; (2) the mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) of each regional air pollution disaster risk index in 2019, as predicted by the PCA-GA-BP neural network, were 0.607, 0.317 and 20.3%, respectively; (3) the predicted results were more accurate than those using a PCA-BP neural network, GA-BP neural network, traditional BP neural network, support vector regression (SVR) or extreme gradient boosting (XGBoost), which verified that machine learning could be used as a method of air pollution disaster risk assessment to a considerable extent (Deng G, 2022).

In this study, the concentrations of air pollutants such as SO2, NOx, PM2.5, O3, and PM10 are examined using machine learning techniques. In order to effectively extract features and make decisions, this research analyses the air quality based on various pollutant concentrations during pre-covid and post-covid days. Based on historical air quality data, a machine learning model is created using logistic regression and decision trees to forecast the air quality index. The results of the experiments demonstrate the effectiveness of the proposed model in determining air quality and forecasting future pollution levels (KB, 2020).

In this research, a brand-new hybrid interpretable predictive machine learning model with two innovations is put forth for the prediction of Particulate Matter 2.5. Deep neural networks and a nonlinear auto regressive moving average with exogenous input model are used to build a hybrid model structure first. Second, this hybrid model incorporates automatic feature generation and feature selection processes. The experimental findings show that the model outperforms other models in terms of peak value prediction accuracy and model interpretability. The suggested model demonstrates how historical PM2.5, weather, and season data are used to estimate PM2.5 forecast. The correlation coefficients for the accuracy of the 1, 3, and 6-hour predictions are 0.9870, 0.9332, and 0.8587, respectively. More significantly, the suggested method offers a fresh interpretable machine learning framework for time series data, allowing for the development of accurate

predictive models and the explanation of complicated interdependencies across multimode inputs. (Gu Y, 2022)

In this study, different error-prone methodologies, including R-Squared (R2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) methods, are listed in order to estimate the AQI value for Particulate Matter (PM2.5) m at a specific area of Delhi using a range of data forecasting ways. The suggested method combines the Long-Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) deep learning models to forecast the AQI of the surrounding environment. In order to compare their performances with the proposed hybrid (LSTM-GRU) model, several stand-alone machine learning (ML) and deep learning (DL) models are trained on the same dataset, including LSTM, Linear-Regression (LR), GRU, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). It is discovered that the proposed hybrid model outperforms all others, with MAE values of 36.11 and R2 values of 0.84. (Sarkar N, 2022)

The Environmental Protection Administration of Taiwan collected data over an 11-year period, which was used in this study to construct a number of prediction models (EPA). For predicting the level of the air quality index (AQI), machine learning techniques such adaptive boosting (AdaBoost), artificial neural networks (ANN), random forests, stacking ensembles, and support vector machines (SVM) show promise. The stacking ensemble consistently outperforms AdaBoost and random forest for R2 and RMSE, but AdaBoost performs best for MAE, according to a series of studies employing datasets for three different areas to acquire the greatest prediction performance from each. (Liang YC, 2020)

Researcher has used Data Mining and Machine Learning models to forecast the AQI and classify the AQI into buckets, in this research project. Principal Component, Partial Least Square, Principal Component with Leave One Out CV, Partial Least Square with Leave One Out CV, and Multiple Regression AQI Data of Multiple Indian Cities are the five regression models we used to predict AQI. Based on the value of the AQI, the AQI Index is further divided into six different categories termed buckets: "Good, Satisfactory, Moderate, Poor, Very Poor, and Severe." Researchers have created three classification models, including Multinomial Logistic Regression, K Nearest Neighbor, and K Nearest Neighbors with Repeat CV Classification method, to forecast the AQI bucket. When just taking into account the fifth component from all the models, the PLS model with Leave One Out Cross Validation performed the best at dimension reduction from the

Air Quality Dataset of Different Indian Cities. PLS model was the most accurate and had the lowest RMSE. In terms of accuracy and AUC, the KNN Model with Repeated CV and Tune Length 10 performed the best from Station Wise Data of Indian Cities. (MM, 2021)

In this study, it was anticipated that because ground-level ozone is a secondary pollutant, it may be predicted for 24 hours based on measurements of the weather and main pollutants like nitrogen oxides and volatile organic compounds. Over the course of a year in Delhi, India, we created software to evaluate hourly recordings of 12 air contaminants and 5 weather variables. Eight machine learning algorithms were adjusted, trained, tested, and compared using cross-validation with hourly data for a full year in order to find the best predictive model. The algorithms, ranked by R2 values, were XGBoost (0.61), Random Forest (0.61), K-Nearest Neighbor Regression (0.55), Support Vector Regression (0.48), Decision Trees (0.43), AdaBoost (0.39), and linear regression (0.39). All models improved in their ability to predict events when trained across five years using different seasons, with a highest R2 of 0.75 during the winter. Bidirectional Long Short-Term Memory generated some of the greatest predictions for seasonal training but was the least accurate model for annual training. When trained with full-year data, the XGBoost model was 90% accurate in predicting the correct category out of five air quality index categories 24 hours in advance. In terms of seasonal predictability, winter (97.3%) is far more so than post-monsoon (92.8%), monsoon (90.3%), and summer (88.9%). These findings highlight the value of testing a wide range of machine learning approaches for certain applications and training them on season-specific data sets. (Juarez EK, 2022)

In this study, an innovative logistic regression technique was used to forecast the level of air pollution in a certain local area instead of the naive bayes algorithm. A dataset containing 32516 records was used to implement and evaluate the methods. Through a programming experiment with N=5 iterations for each approach, different types of air pollution were discovered. The confidence interval is 95%, while the cutoff value is 0.05%. About 80% of the G-power test is applied. When compared to Naive Bayes (97.32%), the novel Logistic Regression algorithm (98.26%) has higher accuracy. Conclusion was that, In comparison to the Naive Bayes method, Logistic Regression has the highest accuracy. (Dilliswar Reddy P, 2022)

In this study, we employ cutting-edge deep learning techniques to predict short-term (multi-step-ahead) air quality in a few areas of Delhi, India. Long short-term memory (LSTM) network

models, which also include more contemporary variations like bidirectional-LSTM and encoder-decoder LSTM models, are a component of deep learning techniques. A multivariate time series technique was utilised by the researchers to anticipate air quality over 10 prediction horizons totaling 80 hours and to produce a long-term (one month out) forecast with quantifiable uncertainties. Results demonstrated that despite COVID-19's impact on the air quality during full and partial lockdown periods, the multivariate bidirectional-LSTM model delivers the best forecasts. (Tiwari A, 2021)

In this study focus was on Data set of New Delhi for predicting ambient air pollution and quality using several Machine learning algorithm. Several ML algorithms were widely used such as Random Forese, SVM, regression, classification. (Sinha A., 2021)

The study's primary goal was to forecast PM2.5 concentration using an artificial neural network (ANN). Between 2014 and 2021, Liaocheng's PM2.5 concentrations decreased on an annual basis. When compared to the same times in 2019, Liaocheng's air quality during and after lockdown periods in 2020 was clearly better. The ANN used in the study had an input layer with 11 parameters, an output layer, and a hidden layer with 6 neurons. First, 80% of the data are utilised for training the ANN, and then 10% of the data are used for verification. (He Z, 2022)

In five key areas of the Tamil Nadu city of Chennai, this article examined the impact of shutdown on the air quality index and other contaminants. The monitoring stations' data on air pollutants such PM10, PM2.5, SO2, and NO2 from 2018 to 2019 (the pre-Covid period) and 2020 to 2021 were studied (during-Covid period). The findings showed that PM10 and PM2.5 concentrations decreased by roughly 48% and 39%, respectively. The pollutants SO2 (25%) and NO2 (10%) have also been seen to significantly decrease. The AQI level in Chennai city was also found to be satisfactory to moderate both before and during the lockdown. Adyar (50.38%) saw the greatest drop in AQI, followed by Nungambakkam (44.18%), T-Nagar (40.31%), Anna Nagar (39.98%), and Kilpauk (30.74%). According to the overall research, environmental protection and pollution control are achieved by implementing appropriate regulations at the right time and place. (Sangeetha, 2022)

3. Present Study

3.1 Statement of the problem:

Various research is being done in this field but still results are not accurate. Until now, previous studies have mainly adopted a traditional machine learning method to analyze the air pollution disaster risk. Moreover, these model are not interpretable. The Studies are not able to explain the interpretation between determining factor and their impact on polluted air.

More research is required for the accurate prediction of AQI. Also South Asia is at the epicenter of ambient air pollution. According to the latest World Air Quality Report, 2020, of the top 40 most polluted cities in the world, 37 are in South Asia. **Air pollution in the region is a health hazard and represents the third-highest risk for premature death.** This requires a country-wide, inter-state, and a regional response.

3.2 Significance of Study

Till now all over the world many research have been done in this area but specific to South Asian region no research has been done so far. the significance of this research is, it will develop a hybrid interpretable predictive machine learning model which can be used to predict accurate AQI, and can be used as a strong measurement tool. If successful, this hybrid interpretable machine learning model can be replicated in the other continents. This model can be used by various research organizations and archeological departments.

This model can help us to fight against the poor air quality. which is a big concern for our health and for our wellbeing and maintain the quality of life. this model can be used for a bigger cause which is to save the environment and planet. it can help us to resort the measures which are causing the depreciation of air quality. The scope of study is not limited to a country but to a bigger region.

This study will be a unique study which will help us to study region wise AQI because AQI affects not only one country but AQI also affects the neighboring country. So it is very important to have a region-wise or continent-wise analysis, rather than country-specific, state-specific, city specific.

3.3 Objectives of study

- To detect the quality of air and predict the level of air pollutants in the future.
- To design a hybrid interpretable model for the more accurate prediction of AQI.
- To compare Various predictive machine learning AQI models successfully applied in other continents.
- To study the scope of predictive machine learning model in determining the quality of life index.

3.4 Research Questions

What all are the determinants of poor quality index?

What will be the repercussion of having poor quality index on the quality of human life?

What is the impact of AQI on the growth of country?

Whether the predictive model is successful in knowing the AQI?

Which hybrid model can successfully predict the more accurate AQI?

What is the scope of the hybrid interpretable predictive machine learning model in knowing the air quality of South Asia Region viz other region?

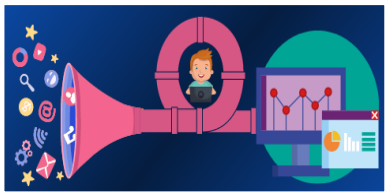
To what extend AQI is a strong determinant of having less development in south Asian region?

4. Methodology

4.1 Research Design

The sequence of steps for the prediction of AQI are explained in this section.

Collection of Raw data



Pre-Processing of data

Proposed Model

AQI Prediction

Evaluate the performance of Proposed Model

4.2 Sample/ Participants

South Asia region - countries of Bangladesh, Bhutan, India, Nepal, and Sri Lanka; Afghanistan and the Maldives are often considered part of South Asia as well.

4.3 Data Collection

The secondary data will be used from the websites like https://aqicn.org/data-platform/covid19/verify/02b17b72-4b2e-4a35-929f-59e739945539, https://datarade.ai/data-categories/air-quality-index, https://www.statista.com/statistics/1122881/apac-pollution-index-score-of-megacities-by-city/,https://www.iqair.com/world-air-quality and https://www.kaggle.com/ . This data would contain the pollutant concentration data (hourly measurement), AQI values along with meteorological data. With the respective information on the breakpoint for each pollutant concentrations.

5. Data Analysis / Discussion

Algorithm which will be used:

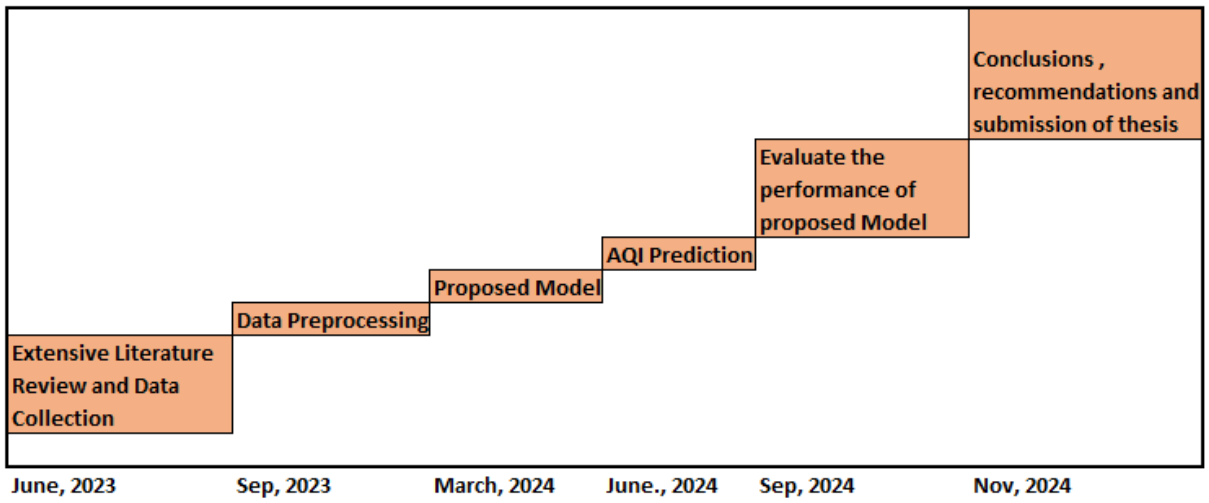
Convolutional Neural Network

A Convolutional neural network has some similarities to the feed-forward neural network, where the connections between units have weights that determine the influence of one unit on another unit. But a CNN has one or more than one convolutional layers that use a convolution operation on the input and then pass the result obtained in the form of output to the next layer. CNN has applications in speech and image processing which is particularly useful in computer vision. Regression model such as

Modular Neural Network

A Modular Neural Network contains a collection of different neural networks that work independently towards obtaining the output with no interaction between them. Each of the different neural networks performs a different sub-task by obtaining unique inputs compared to other networks. The advantage of this modular neural network is that it breaks down a large and complex computational process into smaller components, thus decreasing its complexity while still obtaining the required output.

6. Work Plan of the Research



7. References

Castelli M, C. F. (2020). A Machine Learning Approach to Predict Air Quality in California. *Complexity*, 1-23.

Chen S, K. G. (2018). Investigation of China urban air condition using big data, information theoryand machine learning. *Polish Journal of Environmental Studies*, 1-14.

Deng G, C. H. (2022). Risk Assessment and Prediction of Air Pollution Disasters inFour Chinese Regions. *Sustainability (Switzerland)*, 1-22.

Dilliswar Reddy P, R. P. (2022). Prediction of Air Pollution Level in Particular Region Area Using Logistic Regression and Naive Bayes. *Advances in Parallel Computing Algorithms, Tools and Paradigms*, 618-624.

EG, D. (2010). *Air Quality Index Prediction using K-Nearest Neighbor Technique.* Seria: 103-108.

Gu Y, L. B. (2022). Hybrid interpretable predictive machine learning model for air pollution prediction. *Neurocomputing*, 123-136.

Guerreiro CB, F. V. (2014). Air quality status and trends in Europe. *Atmospheric Environment*, 376-384.

He Z, G. Q. (2022). Prediction of Monthly PM2.5 Concentration in Liaocheng in China Employing Artificial Neural Network. *Atmosphere*, 1-8.

Juarez EK, P. M. (2022). A Comparison of Machine Learning Methods to Forecast Tropospheric Ozone Levels in Delhi. *Atmosphere*, 1-26.

Kang GK, G. J. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. *International Journal of Environmental Science and Development*, 8-16.

KB, I. (2020). ANALYSIS AND PREDICTION OF AIR QUALITY INDEX IN INDIA DURING PRE AND POST COVID PANDEMIC USING MACHINE LEARNING ALGORITHMS. *Palarch's Journal of archaeology of Egypt*, 6995-7003.

Liang YC, M. Y. (2020). Machine learning-based prediction of air quality. *Applied Sciences (Switzerland)*, 1-17.

Madan T, S. S. (2020). Air Quality Prediction using Machine Learning Algorithms-A Review. *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 140-145). Institute of Electrical and Electronics Engineers Inc.

MM, L. (2021). Data Mining and Machine Learning Approach for Air Quality Index Prediction. *International Journal of Engineering and Applied Physics (IJEAP)*, 136-153.

Sangeetha, P. (2022). Assessment of air quality index and pollutants in Chennai city, India: Pre-Covid and during- Covid period. *Proceedings of the International Academy of Ecology and Environmental Sciences*, 269-280.

Sarkar N, G. R. (2022). Air Quality Index prediction using an effective hybrid deep learning model. *Environmental Pollution*, 1-14.

Sinha A., S. S. (2021). Dynamic Forecasing of Air Pollution in Delhi Zone using Machine Learning Algorithm. *Quantum Journal of Engineering, Science and Technology*, 40-53.

Subramaniam S, R. N. (2022). Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review. *Sustainability (Switzerland)*, 269-280.

Sun Y, L. J. (2022). AQI Prediction Based on CEEMDAN-ARMA-LSTM. *Sustainability (Switzerland)*, 1-15.

Tiwari A, G. R. (2021). Delhi air quality prediction using LSTM deep learning models with a focus on COVID-19 lockdown. *Environmental Modelling&Software*, 1-18.